# Efficiently Scaling Transformer Inference

Year: 2022 | Citations: 455 | Authors: Reiner Pope, Sholto Douglas, A. Chowdhery, Jacob Devlin, James Bradbury

---

## Abstract

We study the problem of efficient generative inference for Transformer models, in one of its most challenging settings: large deep models, with tight latency targets and long sequence lengths. Better understanding of the engineering tradeoffs for inference for large Transformer-based models is important as use cases of these models are growing rapidly throughout application areas. We develop a simple analytical model for inference efficiency to select the best multi-dimensional partitioning techniques optimized for TPU v4 slices based on the application requirements. We combine these with a suite of low-level optimizations to achieve a new Pareto frontier on the latency and model FLOPS utilization (MFU) tradeoffs on 500B+ parameter models that outperforms the FasterTransformer suite of benchmarks. We further show that with appropriate partitioning, the lower memory requirements of multiquery attention (i.e. multiple query heads share single key/value head) enables scaling up to 32x larger context lengths. Finally, we achieve a low-batch-size latency of 29ms per token during generation (using int8 weight quantization) and a 76% MFU during large-batch-size processing of input tokens, while supporting a long 2048-token context length on the PaLM 540B parameter model.